Dual-path Adaptation from Image to Video Transformers
-Appendix-

Jungin Park[1*] Jiyoung Lee[2*] Kwanghoon Sohn[1,3†]

[1]Yonsei University [2]NAVER AI Lab [3]Korea Institute of Science and Technology (KIST)

{newrun, khsohn}@yonsei.ac.kr lee.j@navercorp.com

Components	K400 [7]	SSv2 [5]	HMDB51 [8]	Diving-48 [10]
Adapter				
# Adapters per block	4 (2 SP, 2 TP)	5 (2 SP, 3 TP)	4 (2 SP, 2 TP)	4 (2 SP, 2 TP)
Adapter bottleneck width	128	128	128	128
Optimizer (AdamW [15], Cosine scheduler [14])				
Learning rate	3e-4	5e-4	1e-4	3e-4
Weight Decay	5e-2	5e-2	2e-2	3e-2
Batch size	64	128	128	128
Data configuration				
Training crop size	224	224	224	224
Frame sampling rate (T_S)	16 for $T_S = 8$	16 for $T_S = 8$	16 for $T_S = 8$	16 for $T_S = 8$
Frame sampling rate (T_G)	8 for $T_G = 1$ 4 for $T_G = 2$ 2 for $T_G = 3$	Dynamic sampling	8 for $T_G = 1$ 4 for $T_G = 2$ 2 for $T_G = 3$	8 for $T_G = 1$ 4 for $T_G = 2$ 2 for $T_G = 3$
RandAugment [2]	✓	✓	✓	✓
Random erase [17]	✗	✓	✓	✗
Inference configuration				
Testing views (temporal×spatial)	3 ×1	1×3	2×3	1×1

Table A1. Implementation details of DUALPATH.

In this document, we include supplementary materials for “Dual-path Adaptation from Image to Video Transformers”. We first provide more concrete implementation details (Sec. A), and additional experimental results (Sec. B), including the results using a different backbone and ablation study for the resolution of the grid-like frameset. Finally, we visualize more attention maps from each path to complement the effectiveness of the proposed method (Sec. C).

A. Implementation Details

We add parallel adapters in the spatial path and serial adapters in the temporal path to every transformer block. In our adapter, the dimension of the bottlenecked embedding is 128. Following prior work [1], $\mathbf{W}_{\text{down}}$ is initialized with Kaiming Normal [6] and $\mathbf{W}_{\text{up}}$ with zero initialization. For the SSv2 [5] dataset, we additionally insert one adapter before the multi-head attention layer in the temporal path for more robust temporal modeling. The experimental configurations according to the datasets are presented in Tab. A1.

B. Additional Results

B.1. Results with Swin-B

Our DUALPATH can be applied to other transformer-based pretrained image models. We conduct additional experiments with Swin-B [12, 13] transformer pretrained on the ImageNet-21K [3]. The Swin-B contains 24 Swin transformer blocks with 88M parameters, requiring fewer GFLOPs than ViT-B/16 [4]. Each block consists of window-based and shifted window-based self-attention layers. As in the ViT backbones, we add parallel adapters in the spatial path and serial adapters in the temporal path to every Swin transformer block. Note that adapters are attached to only window-based self-attention layers while not adapting shifted window-based self-attention layers. For the SSv2

Method & Arch.	Pretrain	Model # Params	Trainable # Params	GFLOPs	SSv2	HMDB51
Full-tuning w/ Swin-B [13]	IN-21K	88M	88M	124	44.3	61.2
ST-Adapter [16] w/ Swin-B	IN-21K	95M	7M	385	65.1	-
DUALPATH w/ ViT-B/16	CLIP	99M	13M	642	69.3	75.6
DUALPATH w/ ViT-B/16	IN-21K	99M	13M	642	64.7	70.5
DUALPATH w/ Swin-B	IN-21K	97M	11M	287	67.8	75.2

Table A2. Performance comparisons for action recognition on the SSv2 [5] and HMDB51 [8] dataset with different backbones and pretraining datasets.

Method	# Frames	K400 R@1↑	Training GPU Hours ↓	Throughput (V/s) ↑	Inference Latency (ms) ↓
Uniformer-B [9]	32	82.9	5000	3.42	314.58
EVL w/ ViT-B [11]	8	82.9	60	25.53	102.88
DUALPATH w/ ViT-B	16	85.4	31	64.21	15.58

Table A3. Training and inference efficiency comparisons. All models are evaluated using V100-32G, following EVL [11].

Resolution	SSv2		HMDB51	
	GFLOPs	R@1	GFLOPs	R@1
224×224 w/ 16 frames	642	69.3	612	75.6
448×448 w/ 16 frames	846	70.5	816	75.8
896×896 w/ 16 frames	1694	71.6	1632	76.4
224×224 w/ 48 frames	791	71.2	778	76.3

Table A4. Performance comparisons for action recognition on the SSv2 [5] and HMDB51 [8] dataset according to the resolution of the grid-like frameset.

dataset, we use an additional adapter before the multi-head attention layer of the temporal path similar to the ViT backbones. The dimension of the bottlenecked embedding is set to 128.

Tab. A2 provides the experimental results of DUALPATH with Swin-B [12, 13] on the SSv2 [5] and HMDB51 [8] datasets. Although the comparisons between ViT-B/16 and Swin-B backbones show the significantly low computation requirement of the Swin-B model (642 vs 287 GFLOPs with DUALPATH), we attain a comparable performance to the CLIP pretrained ViT-B/16. Compared to ST-Adapter [16] with Swin-B, the results consistently demonstrate the effectiveness of DUALPATH over the backbone networks, showing a higher performance of 2.7% with Swin-B on the SSv2 benchmark.

B.2. Additional efficiency analysis

We additionally compare the methods with [9, 11] in terms of training step time, throughput, and inference latency, following [11]. For a fair comparison, we obtain all results using V100-32G with PyTorch-builtin mixed precision. The throughput is measured with the largest batch size before out-of-memory and the inference latency is measured with a batch size of 1. As shown in Tab. A3, DUALPATH takes about half of the training GPU hours and achieves ×2.5 more throughput and ×6.6 faster inference than EVL [11] under the same hardware condition.

B.3. Resolution of grid-like frameset

The grid-like frameset comprises a stack of 16 *scaled* frames to make the same size as the original frame (224×224). We investigate the effectiveness of the resolution of the grid-like frameset in this section. Note that the impact of scaling factors that determine the temporal resolution is demonstrated in Tab. 5 of the main paper.

Specifically, we set the scaling factors w and h to 1, 2, and 4 while maintaining the temporal resolution as 16 such that the resolution of the grid-like frameset is 896×896, 448×448, and 224×224, respectively. The backbone (ViT-B/16) is identically used and uniformly sampled 8 frames are used in the spatial path. Following [16], we sample one clip cropped into three different spatial views on SSv2 [5] (*i.e.*, total of 3 clips) at test time. For HMDB51 [8], two clips sampled from a video are respectively cropped into three spatial views (*i.e.*, a total of 6 clips). Since a high-resolution frameset contains more detailed information about the original frames, the highest performance is obtained with the 896×896 size of the frameset in Tab. A4. However, the computational cost quadratically increases as the resolution of the grid-like frameset increases. When we use 48 frames (*i.e.*, $T_G = 3$) with the 224×224 size of the frameset, competitive performance is achieved in both datasets. It supports the resolution choice of DUALPATH in terms of the trade-off between performance and computational cost.

C. More Attention Visualization of DUALPATH

The additional attention visualization is illustrated in Fig. A1. We depict the attention maps of $\mathbf{x}_t^{\text{SP}}\{\texttt{[CLS]}\}$ and $\mathbf{x}_g^{\text{TP}}\{\texttt{[CLS]}\}$ from the final transformer block of each path. All videos are sampled from the SSv2 [5] dataset and ViT-B/16 is used as the backbone. While we use 8 frames in the spatial path, the attention maps corresponding to only 4 frames are displayed for visibility. Interestingly, the results show that the model trained with DUALPATH is capable of focusing on dynamic action-related regions in both adaptation paths. As exemplified in Fig. A1a and Fig. A1c, $\mathbf{x}_t^{\text{SP}}\{\texttt{[CLS]}\}$ of the spatial path tends to focus on action-related objects, and $\mathbf{x}_g^{\text{TP}}\{\texttt{[CLS]}\}$ of the temporal path concentrates on action-related movements. Therefore, the two paths complement each other, leading to spatiotemporal modeling.

References

[1] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. In *NeurIPS*, 2022. 1

[2] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with a reduced search space. In *NeurIPS*, 2020. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICML*, 2021. 1

[5] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 1, 2, 3, 4

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. 1

[7] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint: arXiv:1705.06950*, 2017. 1

[8] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, 2011. 1, 2

[9] Kunchang Li, YaliWang, Gao Peng, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatial-temporal representation learning. In *ICLR*, 2021. 2

[10] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *ECCV*, 2018. 1

[11] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *ECCV*, 2022. 2

[12] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, 2022. 1, 2

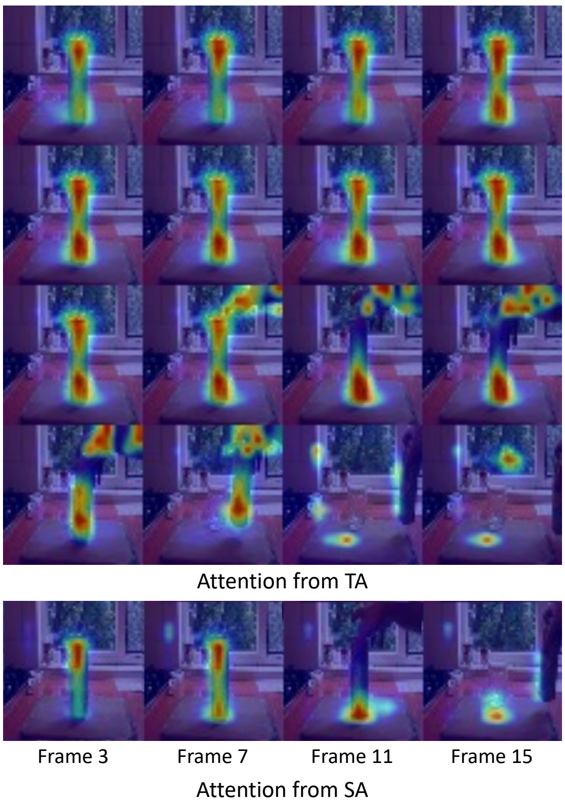
[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1, 2

[14] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 1

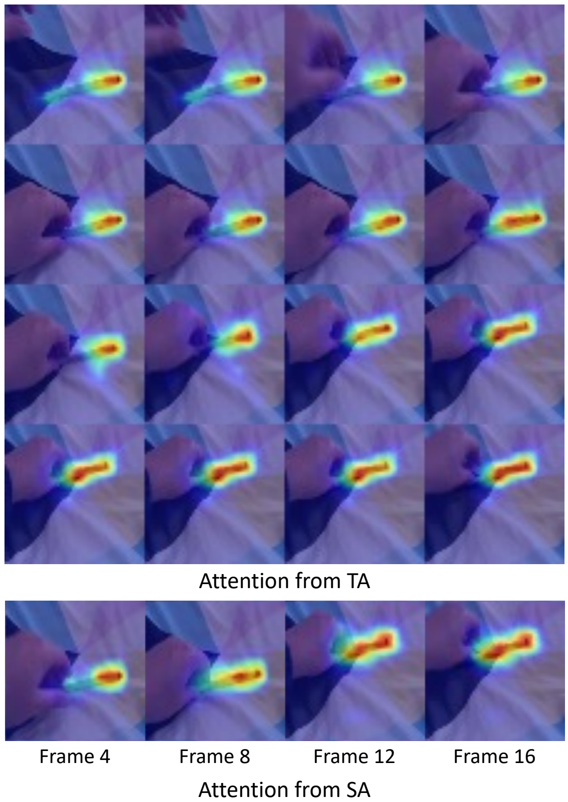
[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1

[16] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning for action recognition. In *NeurIPS*, 2022. 2

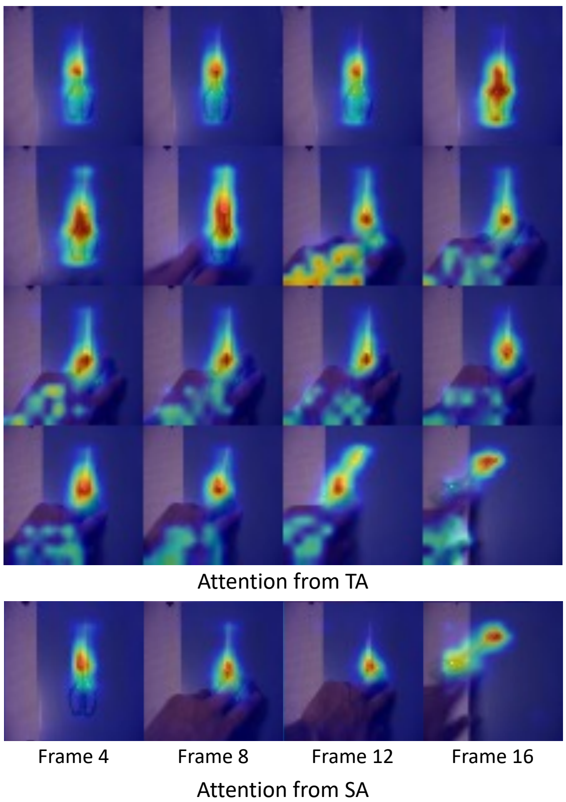
[17] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020. 1



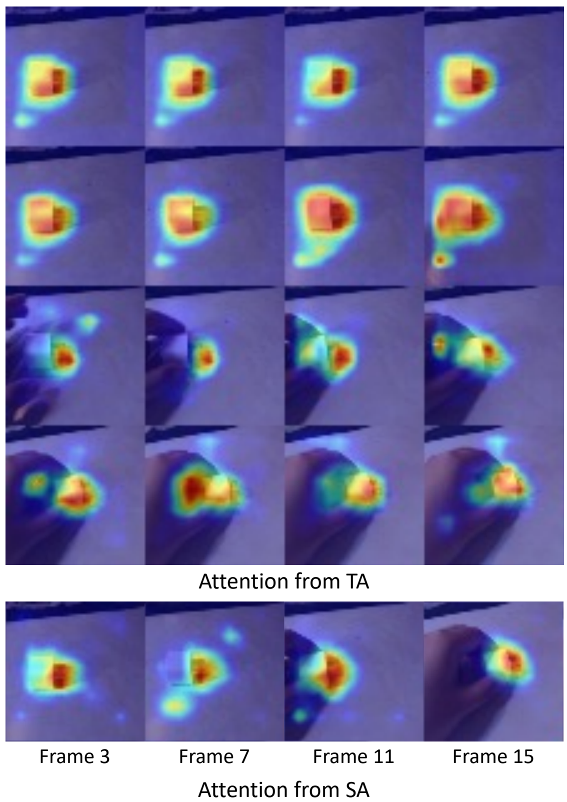
(a) Removing [something], revealing [something] behind



(b) Moving [something] up



(c) Pushing [something] so that it falls off the table



(d) Pushing [something] from left to right

Figure A1. Visualization of attention maps of each path for videos from the SSv2 [5] dataset.